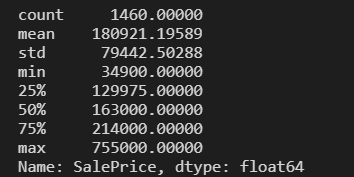
|  |
| --- |
| Harbin Institute of Technology, Shenzhen |
| **Machine Learning Report** |
| House Prices Prediction |
| **Name：张新宇 XinYu Zhang**  **StuId: 19S051009**  **Date: 2019年12月22日** |

## Chapter 1 Data preprocessing

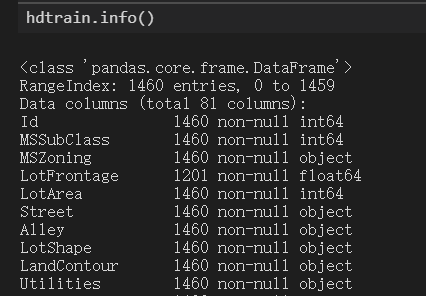
A machine learning project have many things to do, but its goal is clear and easy to specify. For this project (house price prediction), we focused on regression (or prediction) problem and we want to get a good prediction result. And this derived several problems like what is a ‘good’ prediction, how to deal with high dimensional data, how to fetch useful information from data, and how to combine the information to get a good prediction, this is all very important. And we start from the first question.

First step: figure out the indices. From the kaggle website we noticed that they use root mean squared error (RMSE) to evaluate the precision of prediction.

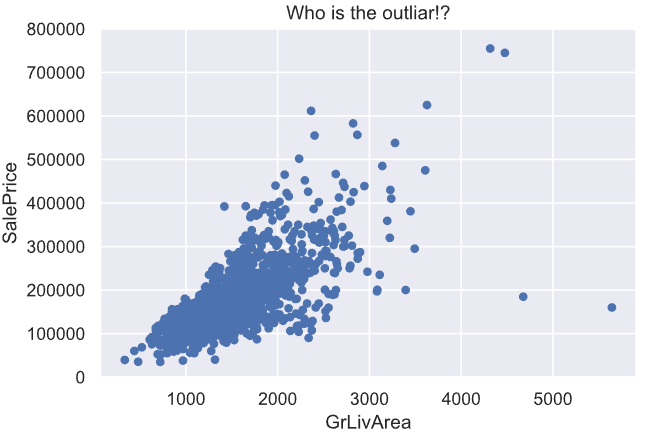
From then we start to deal with the data, we use jupyter to load the data and start analysis the relationship between attribute ‘SalePrice’ ( The prediction we want get) and other attribute in the training data set. But first we should build a basic impression of statistical significance of the ‘SalePrice’ attribute (Use pandas->describe() method)



It is a big difference between max and min, but the mean value and 25%, 50%, 75% is relative. So we may have an assumption about the prices distribution, but it’s not enough. Before we start to analysis the relationships we should figure out the attributes information:



Besides, we can start to plot some graph to build a better understanding about relationships between attribute ‘SalePrice’ and other attributes, first we choose ‘GrLivArea’ to start:



From picture above we can make an assumption that the relationship between two attributes follows the linear distribution, but we see two points in the right that have a big distance between the main points, although it may be reasonable, it still have a big negative influence on our linear assumption. So we decided to remove it from the training set.